9

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek pp. 64–72
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4518

Data from language documentations in research on referential hierarchies

Stefan Schnell
Kiel University

This paper outlines potentials of documentary linguistics for typological research in referential hierarchies. Specifically, I will demonstrate how the analysis of original text data from the Oceanic language Vera'a enhances knowledge about referential hierarchy effects in the domains of number marking and morphosyntactic properties of objects. With this language-specific research as a background, I will outline ways in which original text data from language documentation projects can be used in cross-corpus investigations of aspects of referential hierarchies across languages.

1. INTRODUCTION. This paper[1] outlines potentials of language documentation for typological research in referential hierarchies. After a brief summary of typological grammar-based research in referential hierarchies in Section 2, I will show in Section 3 that certain patterns of number marking and object realization in the Oceanic language Vera'a emerge only through the investigation of original, culture-specific text data. Section 4 outlines how this type of corpus-based research may supplement the established typological approach.

2. TRADITIONAL RESEARCH ON REFERENTIAL HIERARCHIES IN LINGUISTIC TYPOLOGY. Traditional typological research in referential hierarchies has focused on the comparison of languages in terms of the structural variation or restrictions within a specific type of construction. Two classic examples are number marking in referential expressions (Smith-Stark 1974) and the differential realization of arguments in the clause, most notably differential case marking of objects (Bossong 1985). In both construction types, a split is observed between a positive and a negative value for the formal variable in question (presence vs. absence of number marking / case marking, respectively). Positive and negative values are associated with elements in different areas on the Referential Hierarchy (i.e. Silverstein's hierarchy; cf. Silverstein 1976), and the construction split is thus mapped onto

[1] The research reported in this paper was generously sponsored by Grant II/81 898 from the Volkswagen Foundation whom I would like to acknowledge hereby. I am grateful to two anonymous reviewers for valuable comments on an earlier version of this paper. Also, I would like to thank the general editor of this issue of LD&C, Frank Seifart, and the convenors of the Analysis panel of the Leipzig Workshop "Potentials of Language Documentation: Methods, Analyses, and Utilization" Leipzig, 3–4 November 2011, Geoffrey Haig, Nikolaus P. Himmelmann, and Anna Margetts, for further comments and suggestions. I am of course responsible for all remaining errors.

a (cluster of) functional domain(s) comprising person, referentiality (which roughly corresponds to "activation" or "accessibility" (cf. Lambrecht 1994)), and animacy. Figure 1 is a reproduction of this hierarchy rendering the well-known patterns of restricted plural marking in five different languages for different types of referential expression (cf. Croft 2003: 130/134), and it shows two types of distributions that are unattested and indeed precluded. The Referential Hierarchy thus represents a model of possible and impossible linguistic structures or languages (cf. Croft 2003). In analogy, patterns of differential object marking can be mapped onto the Referential Hierarchy; however, additional notions like number distinctions and definiteness have been shown to be relevant here.

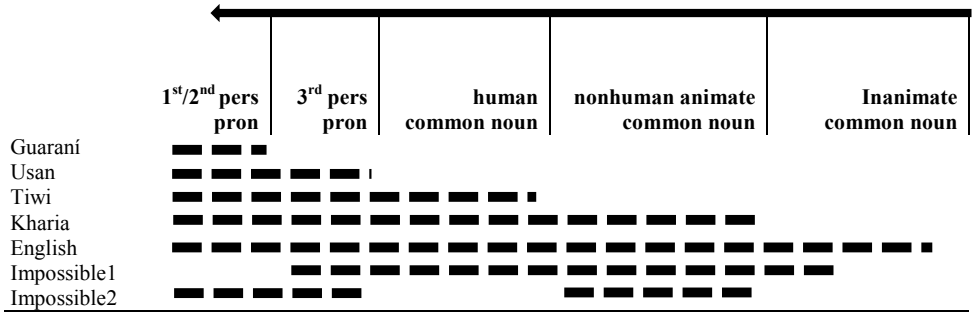

FIGURE 1: Distribution of number distinctions on *Extended Animacy Hierarchy* after Croft (2003: 134)

Crucially, the data bases for this kind of typological work comprise grammatical descriptions or specific studies dedicated to the phenomenon in questions. The latter are often based on focused elicitations of typologically relevant information. The information obtained in this way are interpretations and – to some degree – abstractions of linguistic structures.

Different types of morphosyntactic constructions in individual languages are associated with fixed, rather small feature sets in order to enhance clear distributional descriptions and generalizations. For marking of plurality, only one pair of (usually binary) formal features of a particular element may be considered, i.e. presence vs. absence of plural-marking affix. The values for this variable can then be associated with areas on the Referential Hierarchy (cf. *WALS* Feature #33A, Dryer 2011). Of course, more complex systems may be considered in this way, for instance number systems with more than two values or differential case marking in a P as well as an A function, possibly with more than two possible values (cf. Bickel & Witzlack-Makarevich 2008).

Two problems with this approach remain unresolved in this line of research and can probably only be tackled by use of corpus data: 1. The general neglect of language-internal variation; 2. Treatment of epiphenomenal associations of construction splits with the Referential Hierarchy as connected to factors of discourse structure (cf. Simpson, this volume). Text data has, however, rarely been used for this kind of research (cf. Wälchli 2006), and the purpose of the following sections is to outline how such data can supplement our understanding of animacy and referentiality facts as observable cross-linguistically.

3. INVESTIGATING REFERENTIAL HIERARCHIES IN VERA'A. The Oceanic language Vera'a was documented in a DoBeS project, and the text corpus provided in this documentation served as the main, and almost sole, database for a study in animacy and referentiality effects on the morphosyntax of the language (Schnell 2010)[2]. I will briefly summarize the findings concerning number marking of referential expressions and the differential treatment of P arguments.

3.1. NUMBER MARKING. Possible number distinctions and the means of number marking in Vera'a depend on the type of referential expression and their animacy properties. Pronouns – which most often have human referents (94.9% of pronominal S, A, and P arguments; Schnell cf. 2011b) – show an obligatory 4-way SG-PL-DU-TL/PAUC distinction. Common nouns designating kin relations show an obligatory 3-way SG-PL-DU distinction. Other human nouns designating age- and sex-defined subclasses of humans obligatorily distinguish singular vs. non-singular number and can optionally be marked for dual. With the exception of nouns designating natural forces, all other nouns optionally distinguish singular and plural.

	DISTINCTIONS	MARKING DEVICE
Pronouns	SG-PL-DU-TL/PAUC obligatory	inflection as in Table 2
Kin terms	SG-PL-DU obligatory	+ reduplication pers. DU/PL
'man', 'woman', 'child', … (+hum referent)	SG-NSG oblig; opt. DU	+ reduplication pers. DU/PL
'human being', spirits, animals, inanimate Ns	opt. SG-PL	+/-reduplication PL particle
forces ('hurricane', 'sun', 'fire')	–	–

TABLE 1: Number distinctions and means of marking with different types of nouns in Vera'a

Means of number marking also correlate with referential form and animacy: Personal pronouns are inflected for person and number as shown in Table 2. Kin terms and nouns designating age- and sex-defined subclasses of humans are reduplicated and form a personal NP with *raga* 'people (PL)' or *ruwa* 'two people (DU)' as head noun, as in (1)[3]. All other

[2] Corpus of the Vera'a language compiled by the author is available at http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI649371%23
Another text corpus of Vera'a compiled by Alexandre François can be found at http://lacito.vjf.cnrs.fr/archivage/languages/Vera_a_en.htm

[3] Abbreviations: 1, 2, 3 – 1st, 2nd, 3rd person; A – agent-like argument of canonical transitive verb; ART – article; CS – construct suffix; DEM – demonstrative; DISC – discourse particle; DU – dual; INCL – inclusive;

nouns (except for those designating forces) are preceded by the pluralizing particle *'erē* where non-singular number is to be made explicit, as in (2). Example (2) also demonstrates the optionality of plural marking with the noun *'añsara* 'person, human being' and other common nouns (cf. Table 1).

	SINGULAR	DUAL	TRIAL / PAUCAL	PLURAL
1 INCL	–	*gidu(ō)*	*gidō'ōl*	*gidē*
1 EXCL	*no*	*kamadu(ō)*	*kamam'ōl*	*kamam*
2	*nik(ē)*	*kumru(ō)*	*kimi'ōl*	*kimi*
3	*di(ē)*	*duru(ō)*	*dir'ōl*	*dir(ē)*

TABLE 2: Vera'a free personal pronouns

(1) [1.PALA.009]

e	***ruwa***	***re-reñe***	*anē*	*duru*	*=m*	*da'ō*	*duruō*
ART	two.people	RED-woman	DEM	3DU	=TAM	care.for	3DU

'The two girls, they (DU) [the parents] looked after them [the girls].'

(2) [HHAK.002]

di	*ga*	*kurkur*	***ēn***	***'erē***	***'añsara***	*di*	*ga*	*kur*	***ēn***
3SG	TAM	RED:devour	ART	PL	person	3SG	TAM	devour	ART

'añsara	*delñe*	*=n*	*Vunu Lava*
person	around	=ART	place.name

'He ate the people, he ate the people around Vanua Lava.'

Referentiality and animacy are both relevant for the variable expression of number, with (almost always human) pronouns making all available distinctions, followed by kin terms and other human nouns. Lesser number distinctions are made with non-human and inanimate nouns. With high-ranking common nouns, more complex means of number marking are employed, and these are obligatory; number marking of lower-ranking expressions is less complex and optional. Furthermore, preliminary observations suggest that the occurrence of the optional pluralizing particle *'erē* depends on certain referential properties of the noun quantified and seems to be more likely with human nouns than with other animate or inanimate ones. Crucially, the likelihood of *'erē* occurring with different nouns under different contextual conditions could hardly be determined on the basis of elicitations or isolated examples, but instead requires quantitative investigations of text data.

INTERJ – interjection; LIG – ligature; LOC – locative; P – patient-like argument of canonical transitive verb;PAUC – paucal; PL – plural; POSS.FOOD – classifier food possession; POSS.VES – classifier vessel possession; RED – reduplication; S – single argument of canonical intransitive verb, SG – singular; TAM – tense aspect mood; TL – trial

3.2. P ARGUMENTS. The realization of P arguments is another area of Vera'a morphosyntax where referential hierarchies are relevant. A more comprehensive treatment of P realization is provided in (Schnell 2011a: 34 ff.) and Schnell (Forthcoming), and I will confine myself here to non-lexical topical P arguments. Such topical P arguments are either realized as pronouns within the verb complex ('Pro'), or left implicit ('zero'). The former choice is largely restricted to human, while the latter is preferred with non-human discourse participants, as shown in examples (3) and (4):

(3) Pronominal P argument, human referent (in 4c) [JJQ.120–123]

a. *'ō* *ko-n* *e* *$_i$Qo* *=m* *sal* *[...]*
INTERJ POSS.VES-CS ART Qo =TAM float

b. *ei* *'aluwō* *$_k$dē* *=k* *da* *mē* *$_i$diē*
INTERJ tomorrow 1PL.INCL =TAM do DAT 3SG

c. *$_k$dē* *=k* *van* *'ō* *$_i$di* *m̄e* *=n* *sisidiñ*
1PL.INCL =TAM go carry 3SG DAT =ART bird.catching
'Oh, Qo's canoe is floating. [...] Hey, tomorrow we will do [the following] to him: we will go with him catching birds.'

(4) Zero P argument, inanimate referent [ISAM.005–006]

a. *$_i$[dir]^A* ***=ēk*** ***bigbig*** *$_j$[ēn* *gorē* *=n* *vovoñodo]^P*
3PL =TAM RED:eat ART POSS.food-3PL =ART RED:fish

b. *$_i$[dir]^A* ***=ēk*** ***mul*** ***'ō*** ***kal$_j$Ø^P*** *lē* *=n* *vono-re*
3PL =TAM go carry up LOC =ART home-3PL

c. *$_i$[dir]^A* ***=ēk*** ***big*** *$_j$Ø^P*
3PL =TAM eat
'Then they eat their catch, the take (it) up [the shore] to their village and have (it).'

The correlation between referential form and animacy features of P arguments is, however, merely a soft constraint which is reflected in a strong tendency and may be violated to some degree. Table 3 gives the combined scores of these correlations for three narrative texts (Texts IDs: ISAM, JJQ, PALA)[4]. Hence, contrary to the general tendency, human Ps may occur as zeros (cf. (5)) and non-human Ps as pronouns (cf. (6))[5]:

(5) Zero P argument, human referent [JJQ.200]

dir *=m* *bol* *ēn* *gunu-m* *dir* *man* *row* *'ō'*
3PL =TAM steal ART spouse-2SG 3PL TAM flee carry
'They stole your wife and fled with (her).'

[4] Recorded texts with annotation in ELAN are available in the Vera'a language corpus (under narrative texts) at http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI649371%23

[5] **Bold face font** marks those constituents that are analyzed as VC-internal constituents.

(6) Pronominal P argument, inanimate referent [JJQ.165]
dir ***=m*** ***var*** *ēn* *'ekē* *anē* *dir* ***=k*** ***var*** ***diē*** *di*
3PL =TAM stump ART place DEM 3PL =TAM stump 3SG 3SG
ne ***ōn*** ***'abilin***
TAM lie askew
'They stumped this place, and as they stumped it, it lay askew.'

	PRO	% OF PRO	ZERO	% OF ZERO	TOTALS	% OF
+hum	67	97.1%	6	9.5%	**73**	**28.9%**
% OF +HUM	94.5%		5.5%		**100.00%**	
-HUM	2	2.9%	57	90.5%	**59**	**71.1%**
% OF -HUM	3.4%		96.6%		**100.00%**	
TOTALS	**69**	**100.00%**	**63**	**100.00%**	**132**	**100.00%**

TABLE 3: Humanness and referential form of topical P arguments in Vera'a[6]

The observation that human P arguments are preferably granted pronominal realization while non-human participants are left implicit can only be verified once a sufficiently large amount of text data is investigated. Isolated examples alone, like the ones cited above, would only be suggestive but never decisive (cf. Stoll & Bickel, this volume, for a nuanced statistical treatment of variation in referentiality in Chintang). A further point worth mentioning here is the preliminary observation that this pattern looks slightly different in non-narrative texts where non-human participants with a P function appear to be more readily pronominalized. Hence, the pattern observed for narrative texts may well be an artifact of this particular text type, and 'discourse topicality' may be the real issue. Future investigation of these text data will show whether this observation is borne out.

4. CROSS-CORPUS RESEARCH IN REFERENTIAL HIERARCHIES – THE GRAID INITIATIVE. Given that investigations of original text data in individual languages may contribute enormously to our understanding of referentiality and animacy, the investigation of such text data *across* languages seems to be the most obvious and most urgent thing to do; all the more as large-scale language documentation projects around the world have produced unprecedentedly large amounts of original text data that are easily accessible for linguists.

There appear to be two main obstacles preventing linguists from directly comparing original texts across languages in order to scrutinize the effects of animacy and referentiality across languages (cf. Wälchli 2006: 1). The need for annotating corpora for the relevant features and the enormous workload involved therein (cf. Schultze-Berndt 2006: 2). The need for text data to be minimally comparable. In order to overcome the problem of comparability, researchers have used either parallel texts, i.e. translational equivalents like the *Declaration of Human Rights* or (parts of) the *Bible* (Wälchli 2009, cf. Cysouw & Wälchli

[6] Spreadsheets containing the scores cited here available at http://vc.uni-bamberg.de/moodle/course/view.php?id=9488

2007 for an overview of available parallel texts), or 'content-equivalent' texts elicited with the help of stimuli like the Pear Film (Chafe 1980) or the *Frog Story* picture book (Mayer 1994[1969]). As for parallel texts, although these have been proven a useful database in research within, for instance, lexical typology (cf. Wälchli 2009, 2006), they may not be suitable for research on referentiality and animacy. This is at least suggested by observations from *Bible* translations: As (Mosel & Hovdhaugen 1992: 10f.) show, 'Biblical' Samoan texts differ from texts of indigenous registers in that the former show an unnaturally high proportion of pronominal reference, while the latter prefer zero anaphora. *Pear Stories* have been shown to be a suitable database for research in referential density (Bickel 2003, Stoll & Bickel 2009). *Frog Stories*, on the other hand, have been shown to feature an unnaturally high referential density and do not seem to be amenable for cross-corpus research in referential hierarchies (Foley 2003).

Despite the obvious advantages of and need for controlled data, the comprehensive text corpora compiled in language documentation projects comprise data of the highest quality in terms of authenticity and cultural embeddedness. The GRAID initiative (Haig & Schnell 2011, Haig et al. 2011) touches on this potential by applying a cross-linguistically applicable and easily practicable set of glossing conventions to texts from language documentation projects. GRAID glosses register the referential form, animacy features, and grammatical function of (mainly core) arguments. Hence, once texts are coded in this way, questions like the one concerning the pronominality of P arguments can be tackle d in an immediate and detailed manner, yielding exact figures about correlations between animacy, referential form, and syntactic function. In this way, texts from different languages can be analyzed quantitatively and – at least to some extent – compared in terms of referentiality and animacy. Haig et al. (2011) demonstrate that the original text data they use for their study of pronominal reference shows a surprisingly high degree of uniformity, suggesting that the lack of control for content may actually be of lesser relevance than would be expected. Thus, while Cysouw & Wälchli (2007: 98) state that traditional typology is fruitfully supplemented by parallel-text typology, the study of original texts in linguistic typology may likewise be a worthwhile enterprise (cf. Wälchli 2006). The GRAID initiative opens up such opportunities in the area of referential hierarchy research.

REFERENCES

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.

Bickel, Balthasar & Alena Witzlack-Makarevich. 2008. Referential scales and case alignment: reviewing the typological evidence. In Marc Richards & Andrej L. Malchukov (eds.), *Scales. Linguistische Arbeitsberichte. ARBEITS BERICHTE 86*, 1–37. Leipzig: Universität Leipzig.

Bossong, Georg. 1985. *Empirische Universalienforschung: differenzielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Narr.

Chafe, Wallace L. 1980. The deployment of consciousness in the production of a narrative. In Wallace L. Chafe (ed.), *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*, 9–50. Norwood, NJ: Ablex.

Croft, William. 2003. *Typology and Universals. 2nd edn*. Oxford: Oxford University Press.

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung – STUF* 60(2). 95–99.

Dryer, Matthew S. 2011. Coding of nominal plurality, feature 33A. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library. http://wals.info/feature/33A (03 March, 2012).

Foley, William A. 2003. Genre, register and language documentation in literate and preliterate communities. In Peter K. Austin (ed.), *Language Documentation and Description 1*, 85–98. London: School of Oriental and African Studies.

Haig, Geoffrey & Stefan Schnell. 2011. Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Introduction and guidelines for annotators. Version 6.0. http://vc.uni-bamberg.de/moodle/course/view.php?id=9488.

Haig, Geoffrey, Stefan Schnell & Claudia Wegener. 2011. Comparing corpora from endangered language projects: Explorations in typology with original texts. In Geoffrey Haig, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.), *Documenting Endangered Languages: Achievements and Perspectives*, 55–86. Berlin: Mouton de Gruyter.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Mayer, Mercer. 1994[1969]. *Frog, where are you?* New York: Dial Books for Young Readers.

Mosel, Ulrike. 1982. The influence of the church missions on the development of Tolai. In Rainer Carle, Martina Heinschke, Peter Pink, Christel Rost & Karen Stadtlander (eds.), *Gava': Studies in Austronesian Languages and Cultures. Dedicated to Hans Kähler*, 155–172. Berlin: Reimer.

Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan Reference Grammar*. Oslo: Scandinavian University Press.

Schnell, Stefan. 2010. *Animacy and referentiality in Vera'a*: Kiel University dissertation.

Schnell, Stefan. 2011a. A grammar of Vera'a, an Oceanic language of North Vanuatu. PhD thesis. Kiel University.

Schnell, Stefan. 2011b. Pronominal reference in Vera'a narrative discourse. Paper presented at the International Workshop on Vanuatu Languages. 20–23 October 2011, A.N.U., Kioloa Coastal Campus.

Schnell, Stefan. Forthcoming. Referential hierarchies in three-participant constructions in Vera'a. In: Eva van Lier (ed.), Referential hierarchies in three-participant constructions. Special issue of Linguistic Discovery, in memory of Anna Siewierska.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*, 213–251. Berlin: Mouton de Gruyter.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In R.M.W. Dixon (ed.), *Grammatical Categories in Australian Languages*, 112–171. Canberra: Australian Institute of Aboriginal Studies.

Simpson, Jane. this volume. Information structure, variation and the Referential Hierarchy.

Smith-Stark, Thomas Cedric. 1974. The plurality split. *Chicago Linguistic Society* 10. 657–661.

Stoll, Sabine & Balthasar Bickel. 2009. How deep are differences in referential density? In Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura & Seyda Özçaliskan (eds.), *Crosslinguistic Approaches to the Psychology of Language: Research in the Tradition of Dan Isaac Slobin*, 543–555. London: Psychology Press.

Stoll, Sabine & Balthasar Bickel. this volume. How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications.

Wälchli, Bernhard. 2006. Descriptive typology, or, the typologist's expanded toolkit. Unpublished ms. http://ling.uni-konstanz.de/pages/home/a20_11/waelchli/waelchli-desctyp.pdf (21 March, 2012).

Wälchli, Bernhard. 2009. *Motion Events in Parallel Texts: A Study in Primary Data Typology*. Unpublished Habilitationsschrift, University of Bern.

Witzlack-Makarevich, Alena. 2011. *Typological variation in grammatical relations*. Leipzig: University of Leipzig dissertation. http://www.uni-leipzig.de/~witzlack/Witzlack2010Typological.pdf (21 March, 2012).

Stefan Schnell
s.schnell@latrobe.edu.au